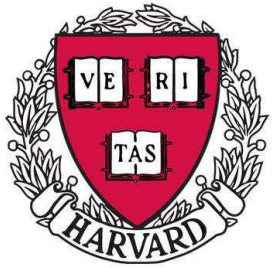
Joint source-filter modeling using flexible basis functions

Daryush D. Mehta, Daniel Rudoy, Patrick J. Wolfe
School of Engineering and Applied Sciences, Harvard University

ICASSP
Prague, Czech Republic
May 26, 2011



Where do we fit in this session to address the complex nature of audio?

What is the best model for speech?

- Let machine learn [Jaitly]
- Peripheral auditory system like [Ellis, Lyon]
- Mimic cortical representation [Mesgarani]
- Speech production based [us]

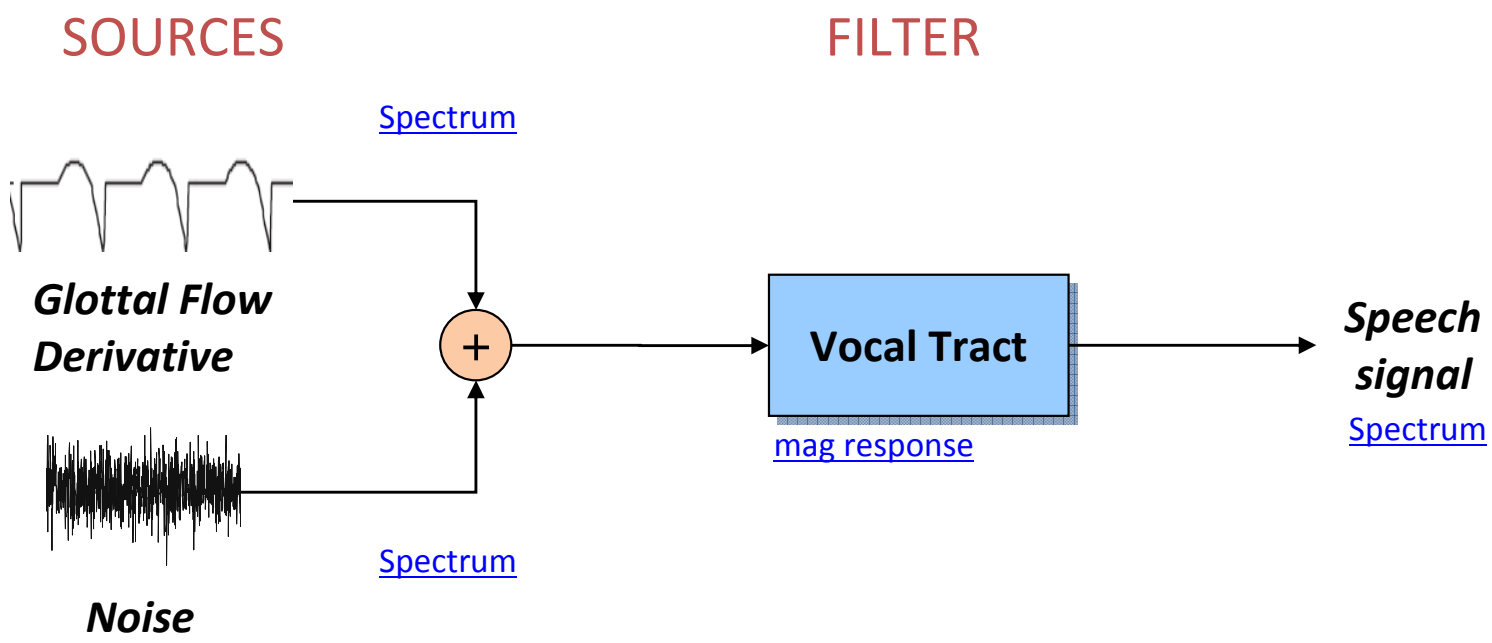
In what subspace do signals lie?

- Spectrogram blocks [Smaragdis]
- Wavelet subspace [us]

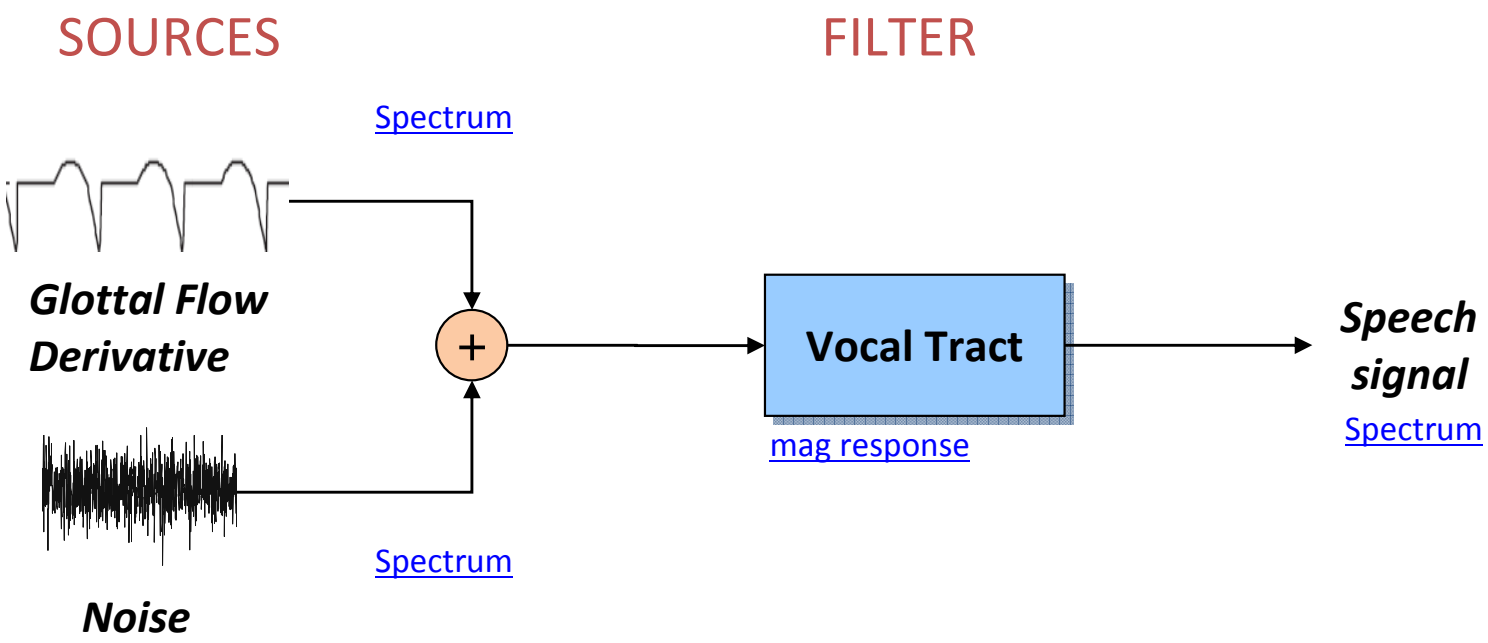
(How do we analyze higher-level information?)

- Texture/soundscapes [Ellis]
- Competing sources [Lyon]

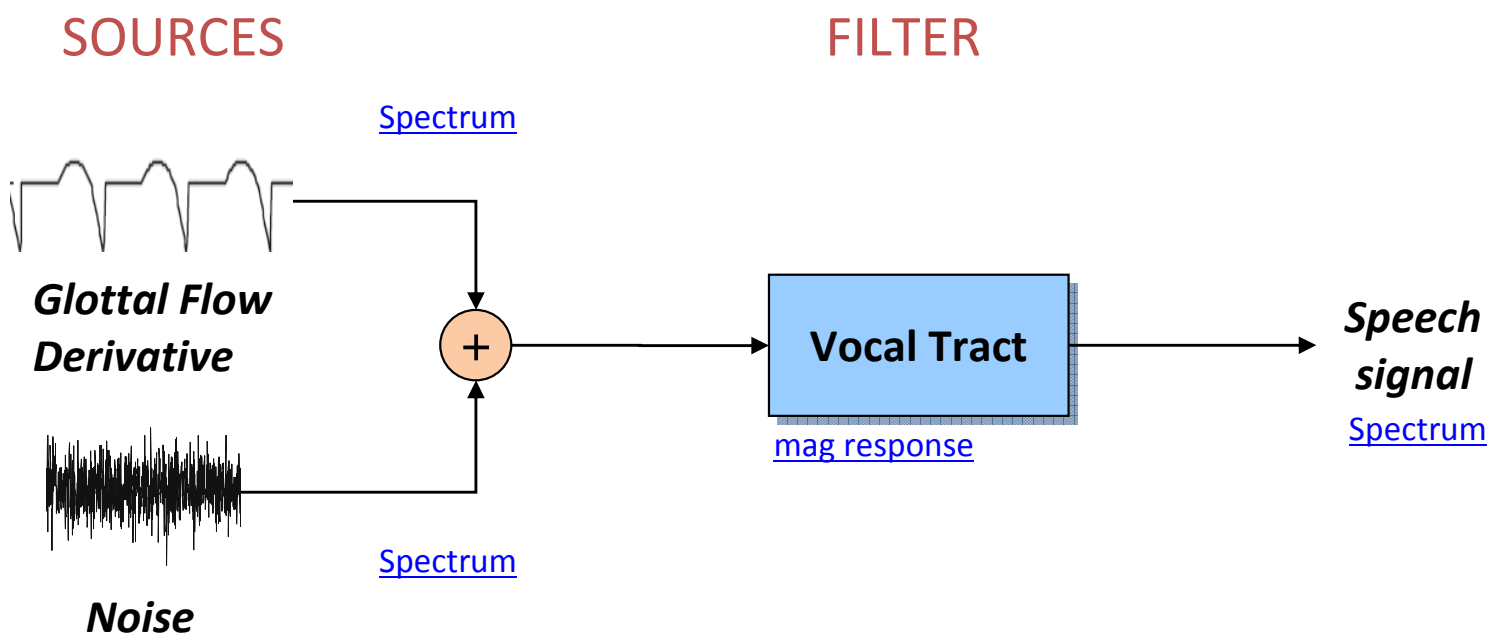
Motivating problem: A model mismatch



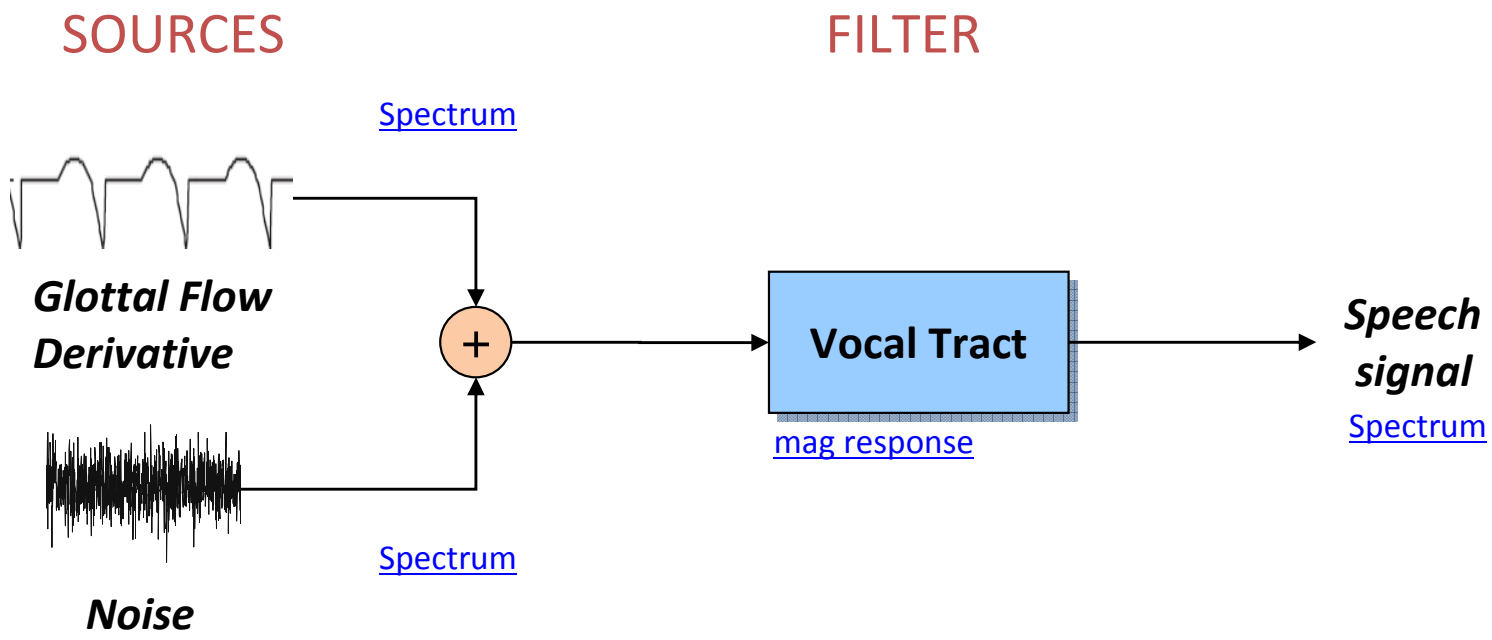
Motivating problem: A model mismatch



Motivating problem: A model mismatch



Motivating problem: A model mismatch

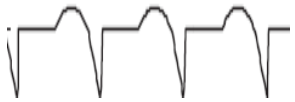


Question: How do we determine appropriate subspace for

1 ?

Background

ARX alternatives to model the non-stochastic voicing component [1]

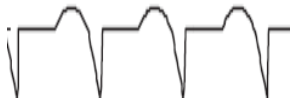


-- Parametric model such as LF model (Vincent 2007) → nonlinear

-- Piecewise linear model (Milenkovic 1986) → glottal opening/closure times needed

Background

ARX alternatives to model the non-stochastic voicing component [1]



-- Parametric model such as LF model (Vincent 2007) → nonlinear
-- Piecewise linear model (Milenkovic 1986) → glottal opening/closure times needed

Our solution is to use wavelets as flexible basis functions (Berezina 2010)

-- Robust to variation in fundamental frequency and irregular pitch
-- Time-localized to preclude need for *a priori* source properties

Background

ARX alternatives to model the non-stochastic voicing component [1]

-- Parametric model such as LF model (Vincent 2007) → nonlinear
-- Piecewise linear model (Milenkovic 1986) → glottal opening/closure times needed

Our solution is to use wavelets as flexible basis functions (Berezina 2010)

-- Robust to variation in fundamental frequency and irregular pitch
-- Time-localized to preclude need for *a priori* source properties

Why do we care?

Clinical voice assessment to inform speech and voice therapy/surgery
(source HNR, spectral tilt, period irregularities, etc.)

Voicing source characteristics important for emotion, health, speaker ID, etc.

Wavelet subspace selection

Least-squares estimators for parameters in ARX model

$$\begin{pmatrix} \widehat{\boldsymbol{a}} \\ \widehat{\boldsymbol{\beta}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{G} \\ \boldsymbol{G}^T\boldsymbol{X} & \boldsymbol{G}^T\boldsymbol{G} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{X}^T \\ \boldsymbol{G}^T \end{pmatrix} \boldsymbol{x}_{N-p}$$

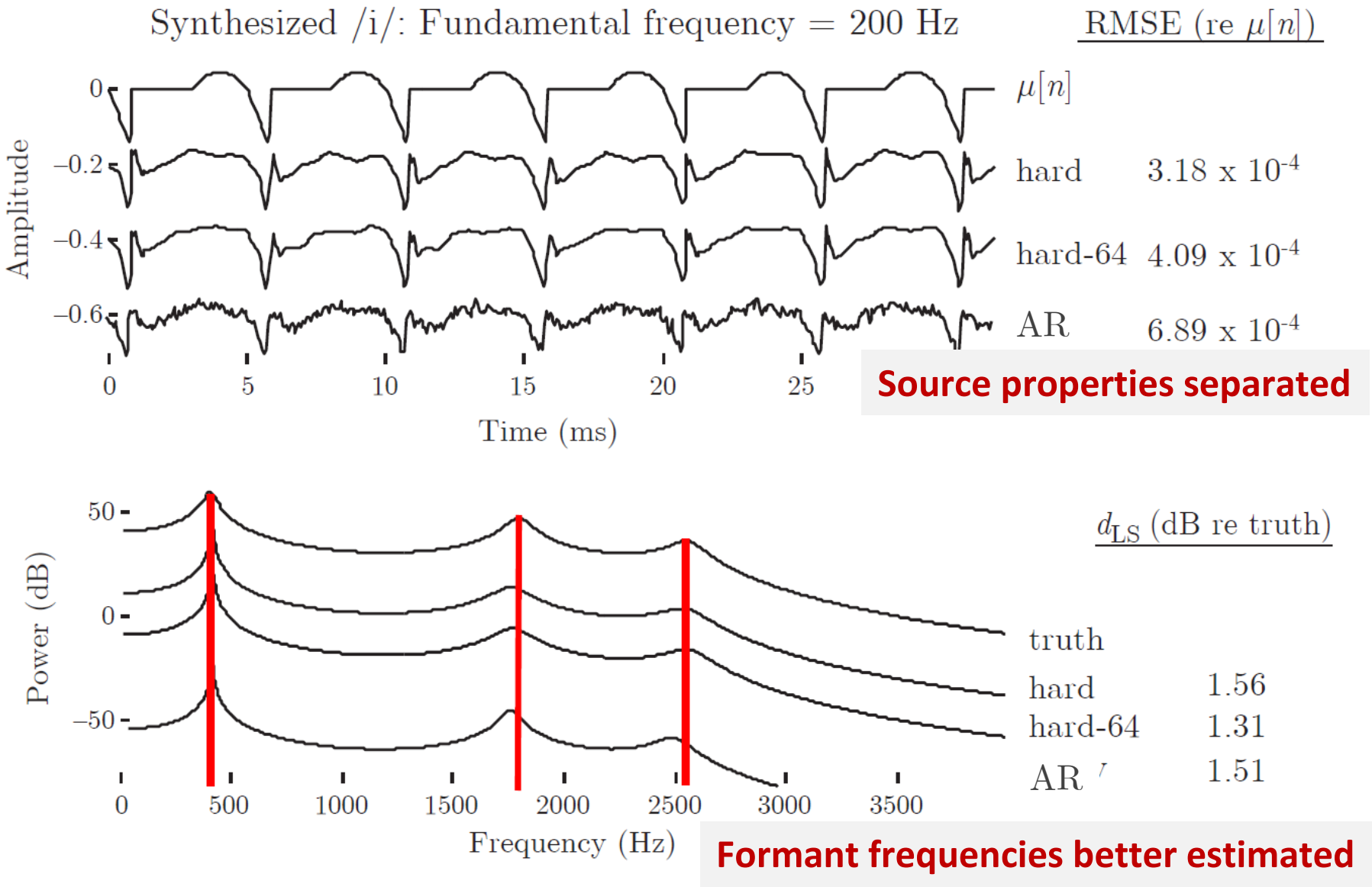
$$\widehat{\sigma}^2 = \frac{1}{N-p} \|\boldsymbol{x}_{N-p} - \boldsymbol{X}\widehat{\boldsymbol{\alpha}} - \boldsymbol{G}\widehat{\boldsymbol{\beta}}\|_2^2$$

Cannot include all wavelets (full rank for) in signal ‘s space
Inversion ill-conditioned

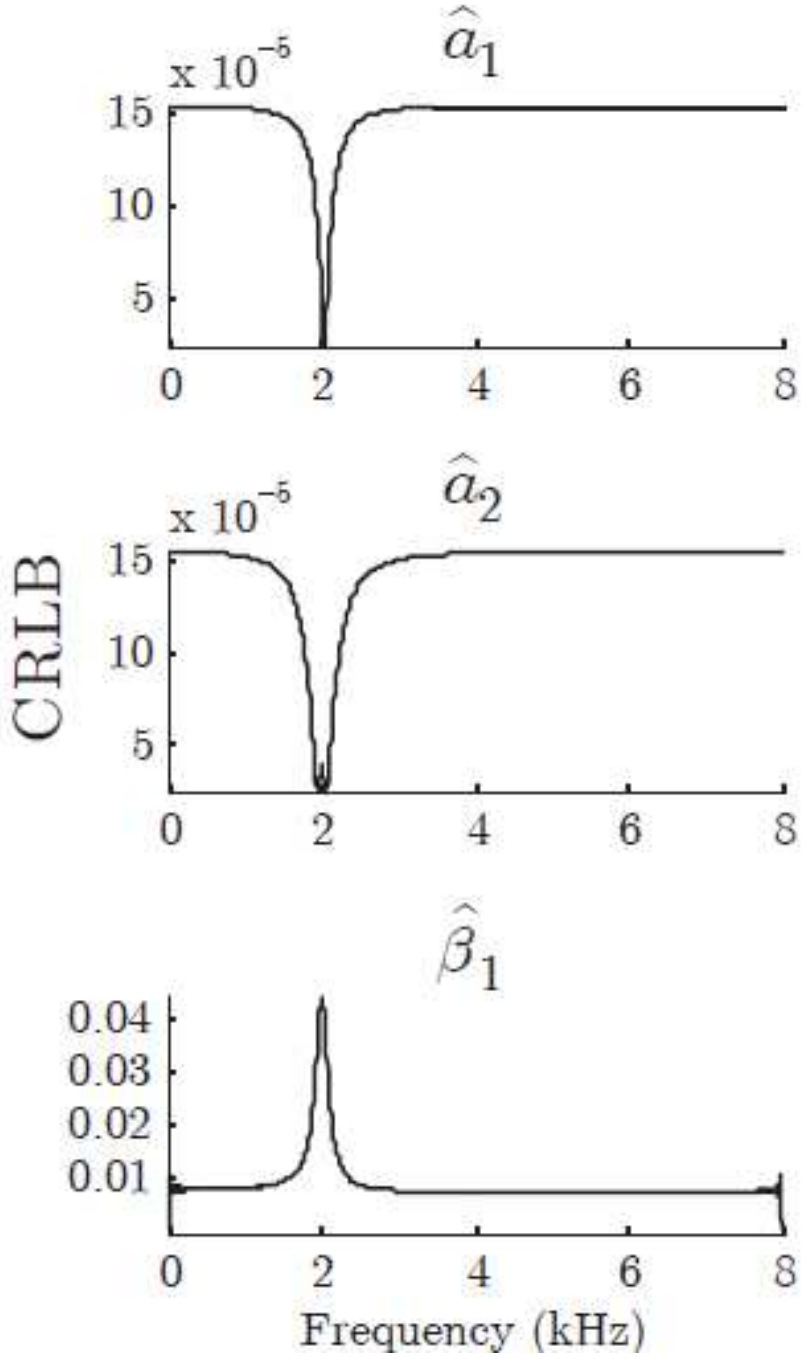
3 wavelet shrinkage algorithms developed

1. Iterative estimation of model parameters and wavelet coefficients
   a. Hard thresholding
   b. Soft thresholding
2. Joint estimation of model parameters using top-N wavelet coefficients
3. Convex optimization of parameters using an $_1$ regularization criterion

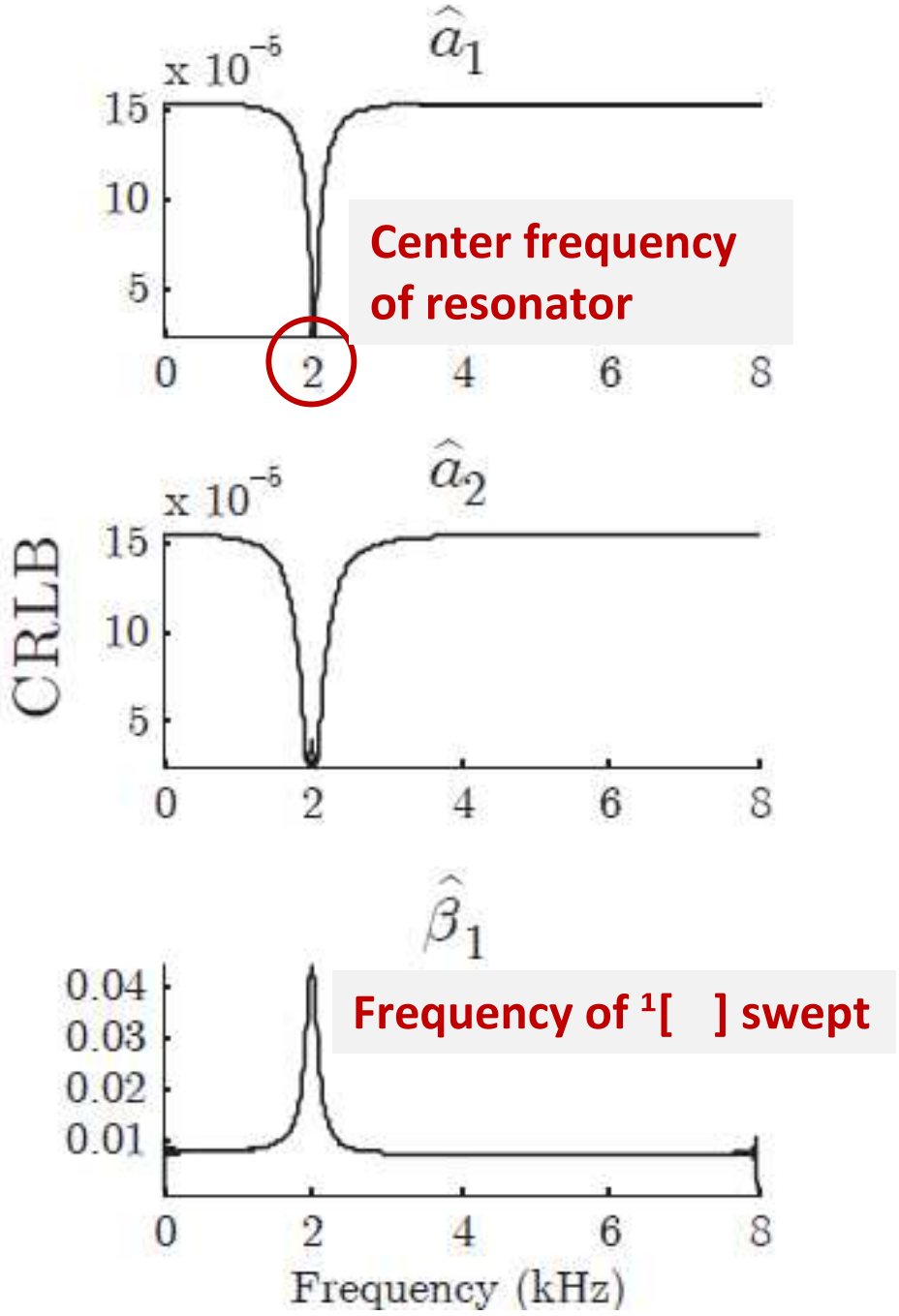
Performance on synthesized vowel



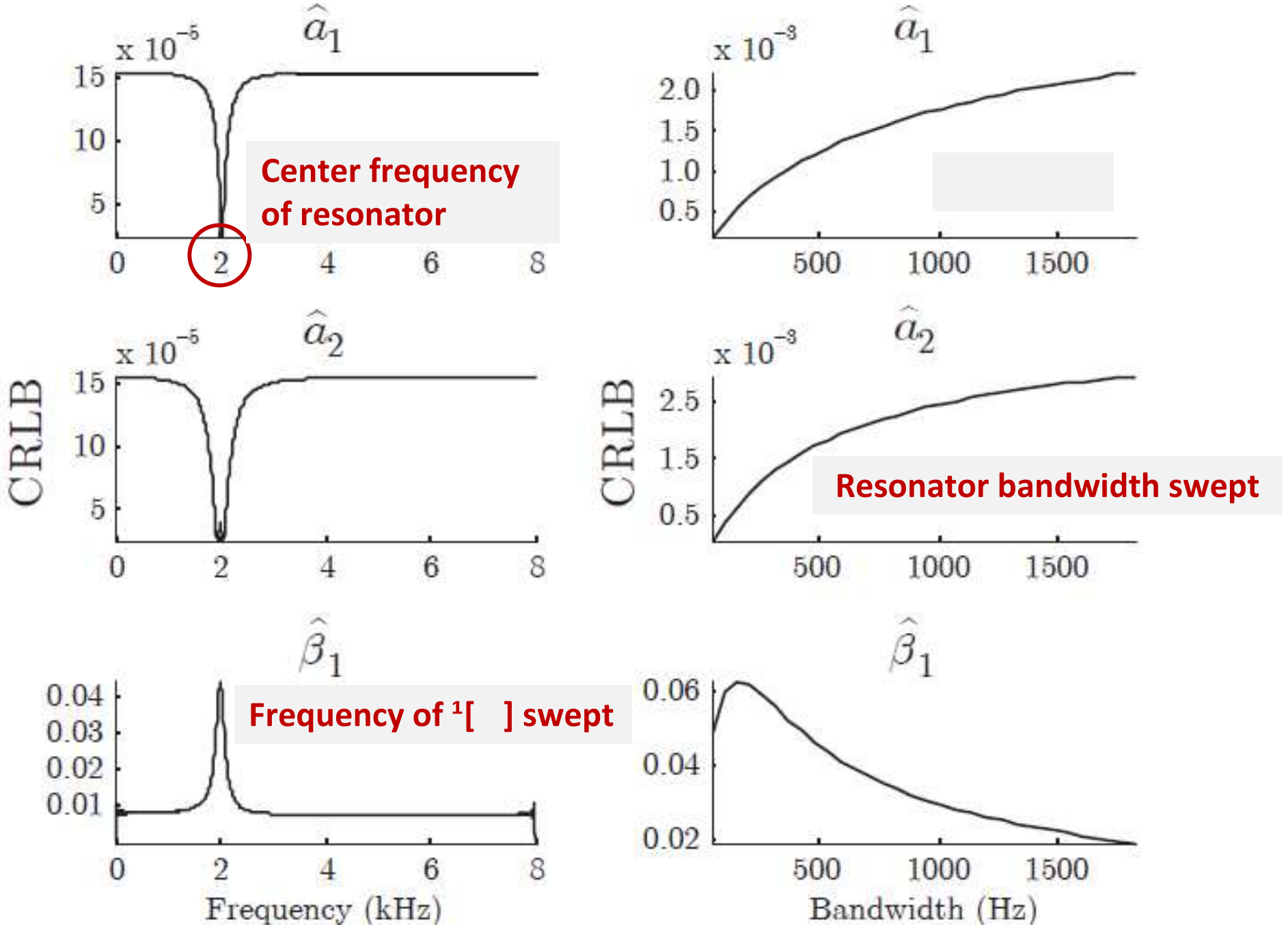
Cramér-Rao lower bounds on estimator variance



Cramér-Rao lower bounds on estimator variance



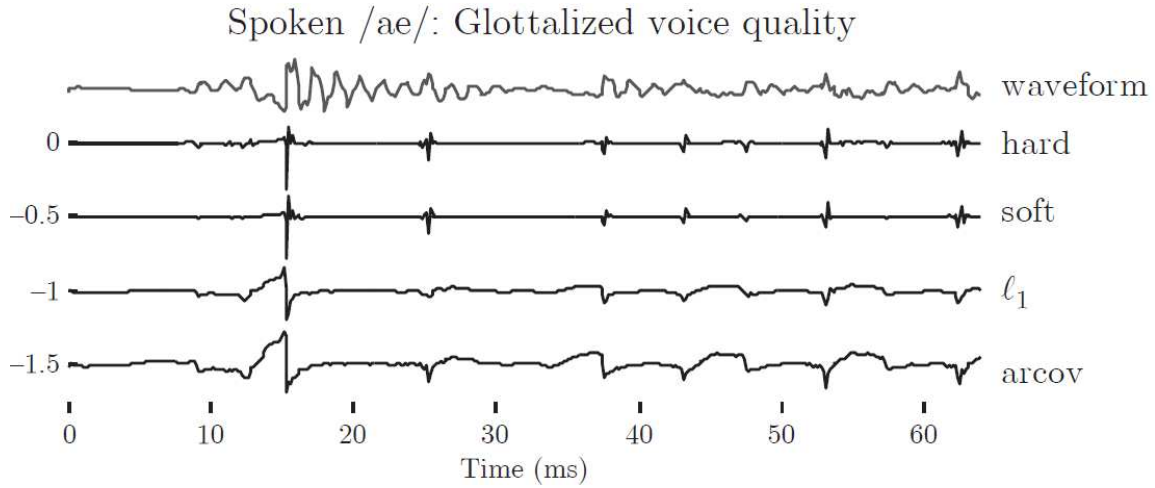
Cramér-Rao lower bounds on estimator variance



Last thoughts

Level-dependent thresholding and penalization of wavelet coefficients to be further explored

Robustness to pitch variation, especially irregular pitch periods, desirable for clinical assessment of disordered voices



Comparison of model outputs with glottal airflow estimation and laryngeal high-speed videoendoscopy measures